Unsupervised Deep Learning for the Discovery of Latent Recurring Structures

A Case Study of the US 30 Index: Technical Summary of Results

John Conaghan
MSc in Artificial Intelligence, University of Limerick
Supervisor: Dr Emil I. Vassev

September 2026

Abstract

This report summarises an empirical investigation into whether unsupervised deep learning (GASF encoding, convolutional autoencoder, clustering) can discover predictive, recurring market regimes from 2.15 million minutes (2018–2023) of US 30 index data. Across pre-registered baselines and ten pre-registered variants, the pipeline recovers volatility states and yields null results for forward-return separation ($\eta^2 \approx -0.00004$; chance-corrected, so values at or below zero mean no effect). Using synthetic regimes planted into the real data as positive controls, the signal is lost at the latent compression stage: variance-driven objectives preserve high-variance volatility structure and discard low-variance trend and directional signals that the GASF encoding itself retains. A six-feature hidden Markov model that recovers planted trend regimes at 10% and 3% prevalence identified no recurring trend or directional state of that kind in the real 2023 market ($\eta^2 = 0.0003$, $p = 0.52$); the null is bounded by the planted family, the one-hour horizon and the covariates removed, and does not extend to regimes those six features cannot see. All confirmatory protocols were pre-registered in version control. The sealed 2024 year was not read by any test in Sections 4–6; the opening-hour test replicated on it under its own rule, and every earlier read is listed in the hold-out ledger.

1 Introduction and problem framing

Research question. Can an unsupervised pipeline discover genuine, recurring market regimes from intraday price action, without access to labels or forward-return targets, and do the discovered regimes carry information about what follows? Put more plainly: does a learned representation of the hour find anything more than volatility?

Answer. It finds volatility states, because volatility is what a variance-driven compression keeps. Every other kind of recurring structure planted into the data was shown, stage by stage, to be lost or unreadable. A simpler method that could see the planted structure found no trend regime of that kind in the 2023 market beyond the clock and the volatility level. The rest of this document is the evidence for those three sentences and the limits on each.

- **Core pipeline.** 60-minute sliding windows of one-minute bars are mapped to four-channel Gramian Angular Summation Field (GASF) images, compressed to a 128-dimensional bottleneck (the *latent vector*, one per window) by a convolutional autoencoder (CAE) trained on reconstruction error alone, clustered by k-means, and only then tested against strictly forward 15-minute returns; the benchmark of Section 6 is tested on the next 60 minutes (Figure 1).
- **Linear yardstick.** Incremental PCA (64 components) runs alongside as the linear compression benchmark.
- **Corpus.** Pepperstone US 30 CFD, January 2018 to December 2023 (∼2.15 M bars). A 60-bar window starts at every minute, giving 2.15 M overlapping windows for fitting; every test is scored on one window per hour, 35,849 non-overlapping hours. Dukascopy provides an independent cross-vendor partition for the same period. Calendar year 2024 is sealed: it may be read only under a pre-registered rule, and every read is logged (see Section 7).
- **Validation metric.** Forward returns are signed log returns measured strictly after each window ends: over the next 15 minutes for the confirmatory tests of the baselines and variants, and over the next 60 minutes for the six-feature HMM benchmark of Section 6, whose outcomes are next-hour trend (lag-1 autocorrelation of one-minute returns) and next-hour direction (signed return). Clusters are compared with a Kruskal–Wallis test summarised by η^2: the share of the spread in forward returns that is accounted

for by which cluster an hour belongs to, on a scale where 0 is none and 1 is all of it. It is chance-corrected (the share random labels would reach by luck is subtracted, so noise scatters around zero and can go slightly negative) and judged against the 95th percentile of a 1,000-draw label-shuffle null. The pre-registered pass mark was 0.06.

- **Sanity check.** Before meeting market data the pipeline was run on ten labelled series from the UCR time-series archive, where the true classes are known, and its clusters were scored against those classes. It recovered them wherever any standard method could (adjusted Rand index, ARI, 0.73 on Coffee and on robot surfaces; ARI scores agreement between two groupings, 0 chance, 1 identical) and scored near zero only where every method fails, so a later null on market data is not a coding fault.
- **Methodological rule.** Every threshold, metric and pass rule was written into a pre-registration file and committed before its test ran. Where a rule was changed afterwards, the change is logged and that result is reported as exploratory rather than confirmatory. Normalisation statistics are fitted on the training years only.

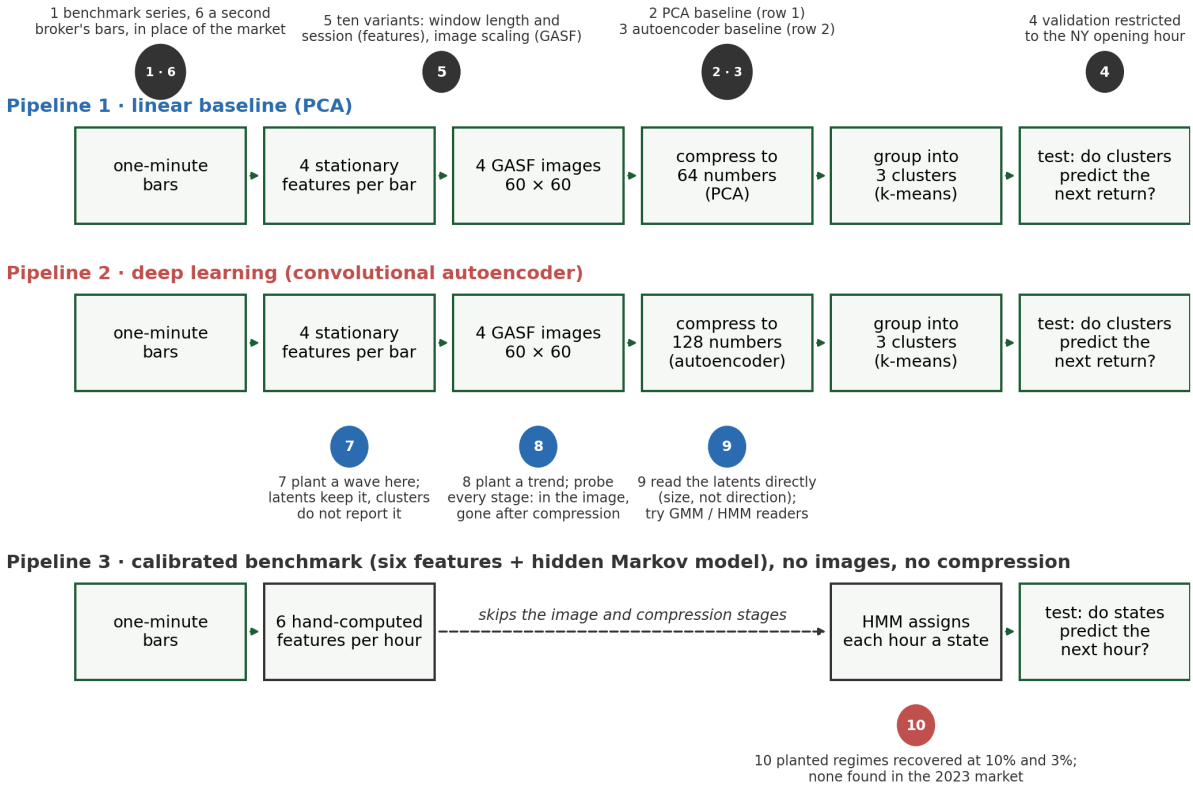

Figure 1: The three pipelines, and where each of the ten experiments acts. Pipelines 1 and 2 share every stage except the compression: 60-minute windows of one-minute bars become four stationary features, then four GASF images, compressed to 64 numbers (PCA) or 128 (autoencoder), grouped by k-means and only then tested against forward returns. Pipeline 3 keeps the bars and skips the images and the compression: six hand-computed features per hour, read by a hidden Markov model. Badge numbers match Table 1: black badges run pipelines 1 and 2 as built on different inputs or settings; blue badges reach inside them, planting regimes and probing the latents; the red badge is pipeline 3.

2 Data representation: four-channel GASF images

Each 60-minute window becomes a $4 \times 60 \times 60$ tensor over four stationary features: log return, normalised range, body ratio and a causal 60-bar volume z-score (labelled price change, bar range, candle body and activity in Figure 2). Values are rescaled to $[-1, 1]$ and mapped to angular space ($\phi = \arccos(\tilde{x})$); pixel $(i, j) = \cos(\phi_i + \phi_j)$, so each pixel is a function of two time points and temporal dependencies appear as image texture. The baselines rescale per window (which discards the hour's amplitude); the size-preserving variant (V1) uses one scale fitted on 2018–2022.

#	Experiment	Question asked	Outcome
1	Benchmark check: the pipeline on ten labelled UCR archive series (ECG, coffee spectra, robot surfaces, leaf outlines, . . .)	Does it recover the known classes where any standard method can?	Yes (e.g. Coffee ARI 0.73, robot surface 0.73, where ARI, the adjusted Rand index, scores agreement between two groupings from 0 for chance to 1 for identical); near zero only where every method fails, so the pipeline code is verified.
2	PCA baseline (Section 3)	Do linearly compressed GASF images cluster into groups that separate forward returns?	Three volatility states; $\eta^2 \approx 0$ on returns.
3	CAE baseline (Section 4)	Does a deep, nonlinear compression do better?	Looser clusters, same states; $\eta^2 \approx 0$.
4	Opening-hour test	Is there structure in the New York open specifically?	Null within sample. Its rule required one replication on 2024: $\eta^2 = 0.016$ ($p = 0.03$), above its own shuffle null but far below the 0.06 bar.
5	Five variant configurations × 2 pipelines (Section 4)	Does keeping amplitude, restricting the session, or changing window length change the answer?	10/10 null; amplitude gives cleaner volatility states only.
6	Cross-vendor refit (Section 3)	Are the clusters a property of the market or of the broker feed?	ARI 0.10 across vendors: partly the feed.
7	Test A, planted wave (Section 5.1)	Is a planted regime kept by the latents, and reported by clustering?	Kept (probe 0.91–0.93); not reported (ARI 0.00).
8	Test B, planted trend (Section 5.2)	At which stage is a variance-neutral regime lost?	Present in the GASF (0.91); gone after compression (0.50).
9	Test C, direct read-out and alternative readers (Section 5.3)	Without clustering, what do the latents hold? Do better readers or a variance-free compression help?	Size yes, direction no; readers recover nothing.
10	Six-feature HMM benchmark (Section 6)	Can a simple model recover planted regimes, and does it find any on the real market?	Recovers at 10% and 3%; finds none in 2023.

Table 1: The experiments, in the order they were run. Rows 2–7 and 10 were pre-registered before they ran; row 1 was a benchmark check without a pre-registration file, row 8 is a post hoc diagnostic, and row 9's reported rule was frozen after a first frozen rule was found mis-specified (Table 2); rows 7–10 were run in September 2026 after rows 2–6 had returned nulls. Exploratory work that is not reported as a result is not listed.

3 Pipeline 1: linear baseline (PCA)

Architecture (row 1 of Figure 1). Flattened four-channel tensors (14,400 inputs) projected onto 64 principal components (incremental PCA), L2-normalised, clustered by k-means; $k \in \{3, 5, 8, 12\}$ chosen by the gap statistic (a rule that compares how tight the clusters are with how tight they would be on structureless data), which selected $k = 3$ in every run.

- **Cluster geometry.** With per-window scaling the clusters are weak (silhouette 0.16; silhouette scores how tight and separated clusters are, 0 for none, 1 for perfect). With amplitude preserved (V1) the clusters are better separated and they last: the silhouette rises to 0.40, and 72% of hours keep their state into the next hour, where 40% would do so by chance (the shuffle null's 95th percentile is 0.40). They also largely coincide with a plain sort of the same hours into low, middle and high thirds of realised volatility (ARI 0.45, where 0 is chance agreement and 1 is identical grouping). The per-window baseline's clusters score 0.00 on that same sort, because its rescaling discards the amplitude before the clustering sees it.
- **Regime identity.** Volatility states aligned with the trading day: *calm* (50% of hours; Asian night; median hourly realised volatility 8 bp, where realised volatility is the square root of the summed squared one-minute returns and an hour's high-to-low range is about 1.5 times it), *thin* (15%; US lunch hour and the hour after the close; 14 bp; fewer ticks than the hour before it, the lowest volume z-score of the three) and *active* (35%; European morning and US cash session; 26 bp). Figure 3.
- **Forward-return separation (2018–2023).** Indistinguishable from noise: $\eta^2 = -0.00005$ (baseline) and -0.00003 (V1) against a shuffle-null 95th percentile of about 0.0001. Under the pre-registered rule a pass on the training years earned one read of the sealed 2024 year; neither run passed, so the baseline pipelines were never evaluated on 2024. A pre-registered test restricted to the New York opening hour returned the same null. Refitting the CAE pipeline on the second vendor's feed reproduced its clusters

Test (Table 1 rows)	Metric	Pass mark, fixed before the run	Outcome
Baselines and variants (2, 3, 5)	chance-corrected η^2 on 15-min forward returns; silhouette; one-lag persistence	all of: $\eta^2 \geq 0.06$ and above the 95th percentile of a 1,000-draw shuffle null (≈ 0.0001); silhouette ≥ 0.20; persistence ≥ 0.50; variants corrected as a family of ten. A pass on 2018–2023 earns one read of sealed 2024.	null: η^2 within 0.0001 of zero in every run and below the shuffle null's 95th percentile; 2024 not read by rows 2, 3 or 5
Test A, planted wave (7)	ARI of k-means labels against planted hours; η^2 on the nudged returns	ARI ≥ 0.20, $\eta^2 \geq$ half the planted target, $p < 0.05$	not recovered: ARI 0.00 at every k
Test B, planted trend, stage-by-stage probe (8)	linear-probe accuracy (chance 0.50)	*not pre-registered*: a post hoc diagnostic run after Test A, reported as exploratory	present in the GASF (0.91); lost after compression (0.48 PCA, 0.50 CAE)
Test C, direct read-out (9)	balanced accuracy of a logistic probe on the frozen latents, 1,000-permutation null	the first frozen rule (tercile accuracy) was found mis-specified at first read and logged as a deviation; the size / direction rule reported here was frozen on 4 September before the second run	size 0.65–0.67 on the V1 latents (0.51–0.52 on the baselines), no better than two volatility numbers (0.67); direction 0.49; planted-wave control 0.62
Stage 15 readers on the compressed latents, 2023 market (9)	chance-corrected η^2 of GMM / HMM states on next-hour endpoints	$\eta^2 \geq 0.005$ in 2023 and the fit years, BH-corrected $p < 0.05$	4 of 24 tests passed, all on the volatility endpoint (η^2 0.055–0.062); a post hoc check attributed about nine tenths of it to a drift in the volatility baseline between fit and score years; reported as exploratory, not as a regime
Pipeline 3, planted regimes (10)	F_1 of the best-matching HMM state against planted hours; planted endpoint	$F_1 \geq 0.50$ *and* the planted endpoint passes the market rule below	recovered at 10% (F_1 0.59) and 3% (0.72); just short at 33% (0.46); negative control clean
Pipeline 3, 2023 market (10)	chance-corrected η^2 of HMM states against next-hour autocorrelation and signed return, clock and volatility regressed out	$\eta^2 \geq 0.005$ in 2023 *and* in the fit years, BH-corrected $p < 0.05$ under a 1,000-draw circular-shift null	bounded null: trend $\eta^2 = 0.0003$ ($p = 0.52$), direction -0.0011

Table 2: Pre-registered pass marks and outcomes. Every rule except Test B's was committed to version control before its test ran; Test B is the one post hoc diagnostic in the main text and is labelled as such.

only weakly (ARI 0.10; an exploratory ablation without the tick-count channel gave 0.33). Reseeding the PCA pipeline on the same feed gives 0.98, so instability under refit is not the explanation, although that check was made on the other model. The clusters are partly a property of the broker's quote engine. The linear pipeline captures volatility scale and offers no directional signal.

4 Pipeline 2: unsupervised deep learning (CAE)

Architecture (row 2 of Figure 1). $4 \times 60 \times 60$ tensor $\rightarrow$ stacked 2-D convolutions $\rightarrow$ 128-d latent bottleneck $\rightarrow$ transposed convolutions $\rightarrow$ reconstruction. Objective: mean-squared reconstruction error; early stopping on 2023 reconstruction loss (disclosed: the CAE has seen 2023 inputs, never returns).

- **Cluster geometry.** The CAE's clusters are diffuse and poorly separated: silhouette 0.06 for the baseline and 0.09 for V1, against 0.16 and 0.40 for the matching PCA runs. Those figures come from each run's

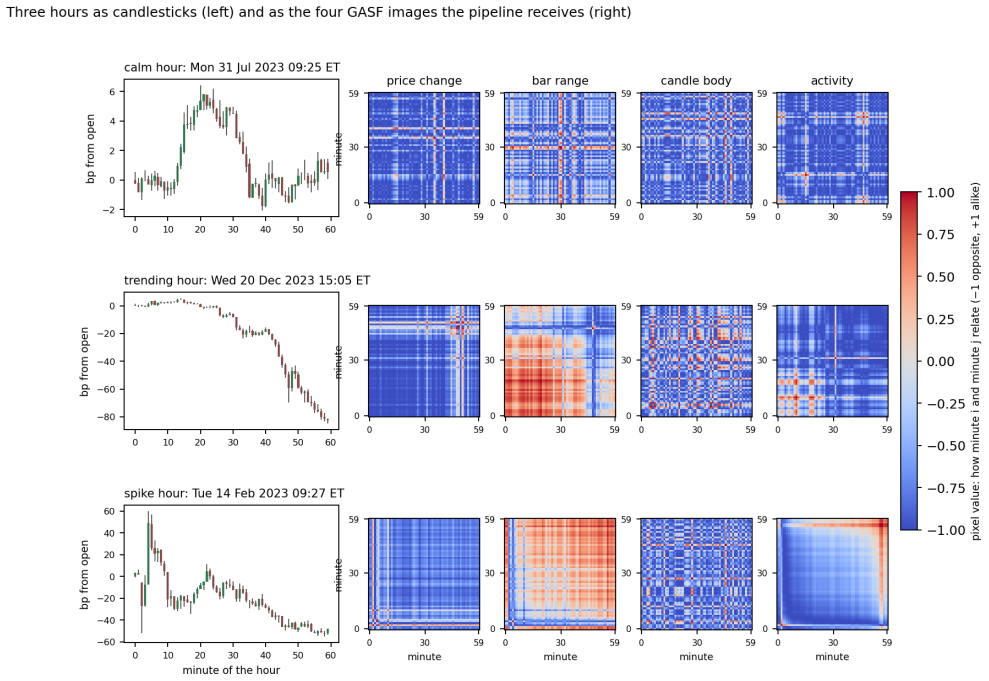

Figure 2: Market hours mapped to GASF textures. Three real hours from 2023, chosen for contrast: calm, trending and a spike at the open. Left: one-minute candlesticks in basis points from the hour's open. Right: the four GASF channels the pipeline receives; both axes are minutes of the hour, and each pixel records how two minutes relate (blue opposite, red alike). The calm hour is mostly one shade; the trending hour shows blocks that flip where the move accelerates; the spike hour is dominated by its first five minutes.

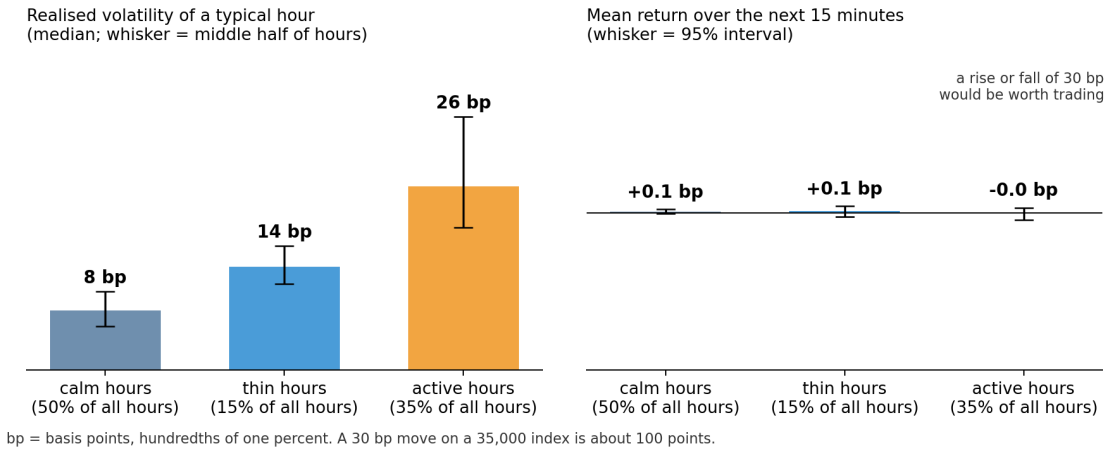

Figure 3: What the clusters are (V1 PCA, amplitude preserved). Left: median realised volatility of the hour itself in each cluster, with the whisker spanning the middle half of hours. The clusters sort hours by how large their bars were. Right: mean return over the following 15 minutes with its 95% interval. All three straddle zero, so the clusters say nothing about the direction of the next move. Whether the size carries forward is not shown here; the direct test in Section 5 answers that (next-move size predicted at 0.65 against 0.50 chance, and no better than two hand-computed volatility numbers).

own sample and its own latent space (64 dimensions for PCA, 128 for the CAE), and silhouette moves with both, so they are not a fair comparison between models. For a fair one, both models were rescored on the same 50,000 windows with both latent spaces reduced to 64 dimensions. The gap stays: CAE 0.067, PCA 0.24. The PCA value falls from 0.40 to 0.24 only because the scoring changed, not the clusters.

- **Regime identity.** The latents sort hours by the same volatility levels as PCA, with weaker boundaries.
- **Forward-return separation.** $\eta^2 = -0.00004$ for the baseline and $+0.00002$ for V1, against a pass mark of 0.06. Five pre-registered configurations (amplitude-preserving scaling, the New York session, its first hour, and window lengths of 30 and 120 minutes) were each run for PCA and for the CAE,

ten tests in all. Because ten tests give ten chances of a lucky pass, their p-values were corrected as a family (Benjamini–Hochberg). All ten are null, and the largest η^2 among them is +0.00004: the clusters account for four thousandths of one percent of the spread in the next move.

- **Interpretation.** The nonlinear pipeline provides no measurable advantage over linear compression for forward-return regime discovery.

5 Diagnostic tests: planted controls

A null result does not distinguish between an absent regime and a pipeline that cannot detect one. Synthetic regimes of known strength were therefore planted into the real inputs as positive controls and followed through the pipeline.

5.1 Test A: retained by the latents, discarded by the clustering

A faint one-cycle wave was planted into the return channel of 33% of hours, with forward returns nudged so that perfect recovery would score $\eta^2 = 0.01$.

- **Linear probe on the latents.** A linear probe is a logistic regression fitted to the latents with the planted labels, asking only whether the information is linearly present. Both PCA (64-d, accuracy 0.91) and CAE (128-d, 0.93) retained the planted regime (chance 0.67).
- **Clustering outcome.** k-means assigned planted and unplanted hours to the same clusters: ARI 0.00 at the pre-registered $k = 3$ (-0.001 PCA, -0.003 CAE) against a recovery bar of 0.20, and at chance descriptively at $k = 2$, 5 and 8. The nudge to the planted hours' forward returns was sized for $\eta^2 = 0.01$ at perfect recovery; the realised oracle value was 0.03, a mis-sizing recorded in the pre-registration's results log that does not affect the ARI verdict. Both representations were refit on the planted corpus, so the result is a property of the method, not of a model that never saw the regime.
- **Mechanism.** The probe is supervised: it is given the planted labels and asked for the direction that separates them, which exists. k-means is given nothing and partitions along the widest gaps in the latent space. The planted regime is present in the latents, but it is a small feature of them: the direction that separates planted hours from the rest carries 2.7% of the latent variance (the unit vector between the planted and unplanted latent means, computed on the full corpus), while the leading principal direction, which tracks volatility, carries 17%, about six times as much; in the autoencoder's latents the shares are 1.1% and 4.8%. k-means minimises within-cluster variance, so it splits along the largest source of variation it can find, and that is volatility. Splitting along the planted direction instead would give a worse value of the k-means objective, so no rerun or reseed would ever choose it.

5.2 Test B: where the planted signal is lost

A second regime was planted, this time into one hour in three and a trending one: each one-minute return was rewritten to carry half of the previous minute's return forward (an AR(1) process with $\phi = 0.5$), so that moves follow through instead of being independent. Each altered hour was then rescaled so that its total variance was unchanged. That step matters because Test A showed the pipeline finds volatility; with the size held fixed, only the shape of the hour marks it as planted. A linear probe, a logistic regression trained to answer "planted hour or not?", was then applied to the data as it stands at each of the four pipeline stages. Where the probe succeeds the trend is still present in the data; where it falls to chance, the trend has been lost (Table 3).

Where in the pipeline	What the probe was given	Probe accuracy	Can the trend still be read?
1. Raw minute bars	The hour's 60 bars as plain numbers (four features per bar)	0.50	No. A trend is a link between one minute and the next, and a linear detector reading bare values cannot see links.
2. GASF image	The return channel of the image	**0.91**	**Yes.** The image turns the links between minutes into pixel values, which a linear detector can read.
3. After PCA	The 64 PCA scores (refit on the 6,000-hour sample; the frozen full-corpus PCA gives 0.50)	0.48	No. Not linearly readable after the compression.
4. After the autoencoder	The 128-number latent vector from the frozen CAE	0.50	No. Not linearly readable after the compression.

Table 3: Where the planted trend is lost. Accuracy is the probe's balanced accuracy at telling planted hours from the rest; 0.50 is chance. Reading the whole four-channel image rather than the return channel alone gives 0.63, because the other three channels carry nothing about the trend and dilute it. The trend survives into the image and is not readable after either compression. "Lost" here means not linearly decodable by the probe; a nonlinear decoder was not tried. For PCA the mechanism is definitional, since it keeps only the directions of largest variance. The autoencoder row uses the frozen model, so it shows what a trained encoder discards, and the attribution to the reconstruction loss is the expected mechanism rather than a measured one.

5.3 Test C: direct read-out and alternative readers

Tests A and B leave an objection open: perhaps the latents do hold a market regime and only the clustering hides it. Test C removes the clustering step and asks the latents directly. A logistic regression was fitted on the frozen 2018–2022 latents, with nothing refitted, and scored on 2023 hours it had never seen. Returns enter only here, in the evaluation, so the learning stays unsupervised. (Two disclosures: the CAE had used 2023 *inputs* to decide when to stop training, and the V1 scaler was fitted on inputs up to the end of 2023; neither saw any return, and the probe's own fit set stops at the end of 2022.) Three questions were asked. Will the next 15-minute move be small or large? On the amplitude-preserving latents the probe answers at balanced accuracy 0.65–0.67 against 0.50 chance (the per-window baselines manage only 0.51–0.52), but two hand-computed numbers, the hour's recent volatility and its volume z-score, answer at 0.67 on the same task, so the latents predict size only through ordinary volatility persistence. Will it be up or down? 0.49, which is chance. Could the probe see direction if it were there? On Test A's planted wave, whose forward returns had been nudged in a fixed direction, the same probe reads direction at 0.62, so the instrument works and the 0.49 is a real absence. The latents carry nothing about direction; the clustering was not hiding a regime, because there was none in the latents to hide.

Two alternative unsupervised readers were then tried in place of k-means: a whitened Gaussian mixture (up to 20 components) and a hidden Markov model. On the compressed latents they recovered no planted regime at 33%, 10% or 3% prevalence. On the real 2023 market the same readers passed the frozen rule on 4 of 24 tests, all on the volatility endpoint (η^2 0.055–0.062); a post hoc check attributed about nine tenths of that to a drift in the volatility baseline between the fit and score years, and it is reported as exploratory rather than as a regime (Table 2). Replacing the compression with 256 whitened directions per channel lets the trend regime survive (probe accuracy 0.82–0.86 at 33% prevalence, 0.73–0.83 at 10% and 0.60–0.70 at 3%, against chance 0.50), and the readers still recovered nothing. The reason is the same as in Test A: a supervised probe is handed the labels and need only find one separating direction, whereas an unsupervised reader must notice the planted hours as a group on its own, and a small shift spread across hundreds of equally weighted directions produces no group it can see.

6 Pipeline 3: calibrated benchmark (six-feature HMM)

Tests A and B showed that the compression step discards a regime that does not change the variance. Pipeline 3 (row 3 of Figure 1) removes that step. Instead of learning a representation, it describes each hour by six numbers chosen by hand for what a trader would look at: how much each one-minute return carries into the next one and the one after (lag-1 and lag-2 autocorrelation), how directly the price travelled from the start of the hour to the end (path efficiency), the hour's realised volatility, the day's realised volatility, and the tick count. Nothing is compressed, so a low-variance signal such as a trend cannot be thrown away before the reader sees it.

The reader is a Gaussian hidden Markov model (HMM). It assumes the market moves between a small

number of hidden states, that each hour belongs to one of them, and that states tend to persist from hour to hour. It was fitted on the 2018–2022 hours. The number of states was chosen from a pre-registered grid of 4, 8 and 12 by the Bayesian information criterion, a standard score that rewards fit and charges for every extra state, so that states are added only when they earn their keep. The market run selected 12, the top of the grid; larger values were not pre-registered and were not tried.

Each 2023 hour was then assigned a state using only the data up to that hour (forward filtering, so there is no look-ahead), and the states were scored on the next hour's trend and direction. Before scoring, the time of day, the day of the week and the volatility level (one-hour, 24-hour and 120-hour realised volatility and the volume z-score) were regressed out of the endpoints by a regression fitted on the fit years, so a state that is merely "the New York open" or "a volatile hour" earns no credit. This is the same fit-and-score split as Test C. The HMM shares no inputs and no parameters with the CAE.

- **Positive control.** The HMM recovered the planted AR(1) trend regime at 10% prevalence ($F_1 = 0.59$) and 3% ($F_1 = 0.72$), and fell just short at 33% ($F_1 = 0.46$), where F_1 scores how well the best-matching HMM state overlaps the planted hours (1 perfect; 0.5 was the pre-registered recovery bar); it stayed at chance on a negative control (a drift on the endpoint with no planted input signature). F_1 is higher at 3% than at 10% because it depends on how the planted hours fall across the model's states: at 10% they were split between two states, at 3% one state captured most of them, so sensitivity is not monotone in prevalence and both figures are reported. This recovery should be read as an upper bound. The planted trend was constructed from the dependence of each minute's return on the previous one, and the HMM's first feature, lag-1 autocorrelation, measures that dependence directly. The model was therefore given the statistic that defines the regime, and its recovery shows what is achievable by a reader matched to the regime, not what an unmatched method would achieve. On the market data the same advantage strengthens the null: a reader matched to a trend regime found none.
- **2023 market.** Each 2023 hour carries a state, and the question is whether hours in different states behave differently over the following hour. Two endpoints were tested. For next-hour trend (minute to minute inside the next hour, do moves follow through or flip?), the states account for $\eta^2 = 0.0003$ of the spread, three hundredths of one percent. The null here is a circular shift rather than a shuffle: because the states persist for hours at a time, the whole state sequence is rotated by a random offset against the outcomes, so the states keep their run lengths but line up with the wrong hours. A value at least this large arises by chance about half the time ($p = 0.52$). For next-hour direction (does the hour as a whole close up or down?), $\eta^2 = -0.0011$, below zero after the chance correction. The pre-registered floor for a regime was 0.005. The planted 3% regime's endpoint effect was written in at a target of $\eta^2 = 0.01$ and read back at 0.012, so the instrument reports an effect of that size when one is present, and the market values are not a failure of the instrument (Figure 4). The floor of 0.005 is a tenth of the baselines' 0.06: the first bar was a conventional medium effect for a first test, while Pipeline 3 was calibrated against a planted effect of 0.01. Both sit far above the shuffle null's 95th percentile of about 0.0001, which every baseline and variant result also fell below, so none of the nulls depends on where the bar was set.
- **Interpretation.** Because the HMM recovered planted regimes at 3% prevalence, its null on the real data can be read as a bound: the 2023 US 30 contained no persistent hour-scale trend regime that was at least as common as the planted ones (3% or 10% of hours) and at least as strong, once the clock and the volatility level are accounted for. A rarer or weaker regime is not excluded: a trend occupying 1% of hours, or one fainter than the planted one, could exist and this test would not have seen it.
  The directional null is narrower. All six features are sign-invariant: autocorrelation, path efficiency, volatility and tick count take the same values whether the hour rose or fell, so the HMM cannot separate up-hours from down-hours on those numbers alone. It can still carry direction indirectly, when a directional regime also changes the shape of the hour in a way the features see. That is what happened with the planted wave: the HMM found the state by its texture (F_1 0.61) and that state's forward returns were tilted, so the endpoint passed. A regime whose only signature is the sign of returns, with nothing else about the hour changed, is outside its reach, and the directional null does not exclude one.
  The CAE null cannot be read as a bound at all. The CAE pipeline never recovered a planted regime at any prevalence, so its market null cannot separate "nothing there" from "could not have seen it".

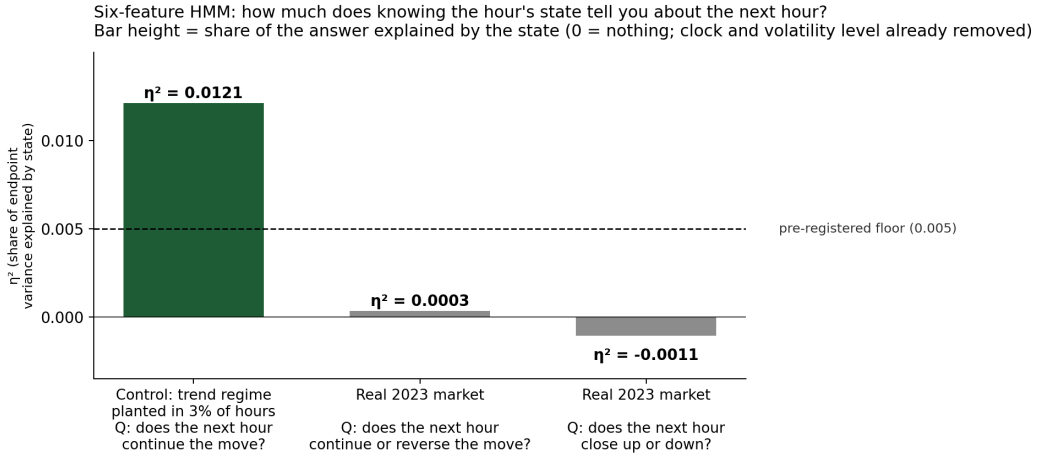

Figure 4: Six-feature HMM on the 2023 market. All three bars come from the same six-feature HMM. The first is its run on data with a planted trend regime, the check that the instrument works. The other two are its run on the real 2023 market, asked two different questions about the hour that follows a state. The first is about behaviour inside that hour, minute to minute: do moves follow through (trend) or flip (mean reversion)? The second is about the hour as a whole: does it close up or down? A trending hour can close either way, so the two are separate questions. η^2 is the share of that next-hour outcome that knowing the state explains, with the clock and the volatility level already removed; values near or below zero mean the states explain nothing, and the dashed floor is the smallest effect agreed in advance to count. The planted regime (green) clears it; the real market (grey) does not on either question.

7 Summary of contributions

Contribution	Empirical finding
1. Where the represent-then-cluster design fails	Planted regimes of known strength showed that variance-driven compression (PCA; MSE-trained autoencoder) discards low-variance temporal structure that the GASF encoding retains, and that two repairs (alternative readers; a variance-free compression) do not recover it.
2. Planted-control auditing methodology	A positive-control framework for unsupervised pipelines: pre-register the pass marks, plant a known regime, verify instrument sensitivity and state the smallest recoverable prevalence before accepting a null. The project's own early nulls fail this standard.
3. Cross-vendor dependence	Latent-cluster agreement across independent broker feeds (Pepperstone vs. Dukascopy) is ARI 0.10 (CAE pipeline), rising to 0.33 in an exploratory ablation without the tick-count channel: regimes found on one CFD feed are not automatically properties of the market.
4. Bounded real-market null	In the 2023 US 30 at the one-hour horizon, no recurring trend regime at the tested prevalences (3% and 10%) and the planted strength ($\phi = 0.5$), and no directional regime of a kind visible to six sign-invariant features, was found beyond the clock and volatility level ($\eta^2 \leq 0.0003$). Calendar year 2024 was not read by any test in Sections 4–6. It was read twelve times between June and July 2026, by the opening-hour, pattern-miner and pre-market replications of the earlier programme, by three Stage 13 tests including the later-withdrawn calm-state confirmation, and by five exploratory checks; two sandbox one-shots also read into 2025–26. Each read is listed in the hold-out ledger in the thesis appendix. The null of Section 6 is therefore bounded to 2023, and whether to spend a further pre-registered confirmation on 2024 before submission is an open decision.

With hindsight

Two of the failure modes follow from textbook properties and could have been tested before the deep model was built: PCA keeps the largest-variance directions by definition and k-means cuts along the widest gaps, and in financial returns the largest variance is volatility; and per-window rescaling of the GASF input discards amplitude. Neither is stated as a limitation in the finance literature this work builds on. The GASF papers apply the rescaling without comment and present the encoding as bijective and information-preserving (Wang and Oates, 2015); the closest imaging-plus-clustering study reports that its clusters capture

volatility and treats that as success (Wu et al., 2023); the regime-clustering literature reports volatility regimes as its intended result (Prakash et al., 2021; Horvath et al., 2021). The amplitude loss is recognised outside finance: an electrical-load study notes that standard GAF images "usually miss" amplitude and adds channels to restore it (Wang et al., 2023). The design was adopted because the papers that use it report positive results under internal metrics, and testing it under external validation was the reasonable first step. What could not have been anticipated from the literature is the measured part: that a planted sub-dominant regime survives the compression yet is discarded by clustering, that a variance-preserving trend regime is lost at the compression itself, and the prevalence at which a simpler method recovers it. The measured results are the contribution; the earlier nulls are the reason they were run.

Alternatives considered

Any representation trained on a reconstruction objective (1-D convolutional, recurrent or transformer autoencoders included) is subject to the same variance dominance shown in Section 5 (Test B): the loss is paid mostly on high-variance volatility content, so the argument, though demonstrated here only for PCA and the CAE, is expected to transfer. Two earlier retrofits of the deep pipeline, one re-injecting amplitude at the CAE bottleneck and one replacing the reconstruction loss with a contrastive objective, are reported in the thesis appendix as attempts whose treatment, on audit, never reached the model (the first was not seen by the loss function; the second was fed unstandardised inputs), so they carry no evidence for or against either idea. A correctly scaled contrastive or persistence-aware encoder would be the next design to test. Computing the properties of interest directly and reading them with an HMM (Pipeline 3) bypasses the bottleneck, and was the only reader here that recovered a planted regime.

References

Horvath, B., Issa, Z. and Muguruza, A. (2021), 'Clustering market regimes using the Wasserstein distance', arXiv preprint arXiv:2110.11848. Accessed 6 September 2026.

Prakash, A., James, N., Menzies, M. and Francis, G. (2021), 'Structural clustering of volatility regimes for dynamic trading strategies', *Applied Mathematical Finance* **28**(3), 236–274.

Wang, J., Wu, Y. and Shu, L. (2023), 'A nonintrusive load identification method based on improved Gramian angular field and ResNet18', *Electronics* **12**(11), 2540.

Wang, Z. and Oates, T. (2015), Imaging time-series to improve classification and imputation, *in* 'Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)', Buenos Aires, pp. 3939–3945. arXiv:1506.00327.

Wu, J., Zhang, Z., Tong, R., Zhou, Y., Hu, Z. and Liu, K. (2023), 'Imaging feature-based clustering of financial time series', *PLOS ONE* **18**(7), e0288836.